Biological Data Mining Book Chapters
19DEC2008

Title: Statistical Approaches to Gene Expression Microarray Data Preprocessing

Author List: Megan Kong[1], Elizabeth McClellan[2,3], Richard H. Scheuermann[1,2], Monnie McGee[3]

Affiliations: [1]Department of Pathology, [2]Division of Biomedical Informatics, University of Texas Southwestern Medical Center, Dallas, TX; [3]Department of Statistical Science, Southern Methodist University, Dallas, TX

Broad Categorization: DNA microarray data analysis

Abstract: Gene expression microarrays have rapidly become a standard experimental tool in the modern biomedical research laboratory. The ability of this technology to interrogate the expression patterns of the entire genome of an organism is fueling systems-level modeling of cells and organisms. The current technologies used for microarray analysis have brought with them analytical challenges related to the inherent noise, variability and quantity of the data generated requiring advanced and novel statistical approaches. In this chapter we will discuss the current statistical methods for probe annotation, background estimation, normalization and summarization, and approaches that can be used to test the effectiveness of these methods on the resulting data.

I. General Overview

A. Gene Expression Microarrays

Over the past decade, gene expression microarray has revolutionized the way we measure the level of gene transcripts in biological organisms. Traditionally, researchers used northern blotting or polymerase chain reaction (PCR) to measure the expression level of a few genes in a single experiment. Now, the expression levels of thousands of genes can be measured in one single microarray hybridization. Researchers can view the expression profiles of genes on a whole genome scale.

Gene expression microarray consists of series of microscopic spots of DNA polynucleotides stabilized onto a slide in a defined array. Each spot corresponds to one polynucleotide sequence. Each of these polynucleotide probes, is designed to be complementary to a specific gene transcript, and is used to capture fluorescently- or radioactively-labeled polynucleotide targets derived from the specific gene transcripts under appropriate hybridization conditions. The relative abundance of target gene transcripts can then be inferred by measuring the fluorescent or radioactive signal associated with each probe.

The gene expression microarray technology has been applied to many areas of basic biological and clinical research. In basic research, using gene expression microarray, researchers can study gene function by examining gene expression pattern differences in samples from wild type and mutant organisms. Gene interaction networks,

which reflect functional relationships among genes, can also be constructed using gene expression microarray data. In clinical research, gene expression microarray has been applied to biomarker discovery by comparing the expression profiles of diseased and normal tissues. It has also been used to study drug response characteristics and determinants of drug resistance. Particularly, by studying the gene expression pattern differences between different study groups, clinicians can gain insight into why some people respond to (or show resistance to) certain drug treatments while others do not.

There are two major public gene expression microarray data repositories: GEO (http://www.ncbi.nlm.nih.gov/geo/) and ArrayExpress (http://www.ebi.ac.uk/microarray-as/ae/). Currently, GEO has 237,664 arrays and ArrayExpress has 181,810 arrays available. These two major data sources have allowed researchers, biostatisticians and computational biologists to compare their results with results from similar studies by other groups, to develop better statistical methods for processing microarray data and to perform data mining to discover new features of the underlying biology.

Four different technology platforms have been used extensively for gene expression microarray studies, each with associated advantages and disadvantages. The initial microarrays were constructed using long cDNA copies of mRNA transcripts as the probes that are spotted on the array. The advantage to using these spotted cDNA arrays is that because the probes are relatively long the stringency of hybridization can be more easily controlled, such that the signal detected is more specific for the transcript targeted. The disadvantage is that the procedures required to synthesize and quality control the probes used is very labor intensive making it difficult to produce these arrays with acceptable reproducibility.

The second technology utilizes arrays in which oligonucleotides (~60mers) are spotted on the array support to serve as hybridization probes. While there is some loss in hybridization specificity, the advantage is that the chemical production of these synthetic oligonucleotides can be controlled more precisely than cDNA synthesis leading to improve reproducibility in probe production. The disadvantage is that spotted oligonucleotide arrays are subject to variability associated with the spotting technique used.

The third technologies utilizes arrays in which oligonucleotides are synthesized directly on the array solid support utilizing photolithography techniques, using custom masks, as in the Affymetix GeneChip™ technique, or digital micromirrors, as in the NimbleGen maskless array synthesis technique, to control light exposure during synthesis. The *in situ* synthesized oligonucleotide arrays offer enhanced reproducibility in array production in comparison to the spotted oligonucleotide arrays.

The fourth technology platform utilizes in situ oligonucleotide synthesis to produce the hybridization probes, but instead of using glass slides uses individual microbeads as solid support for the array. This allows the oligonucleotides to be synthesized using traditional organic chemistry approaches instead of photolithography. Microbead-based oligonucleotide arrays offer many of the same advantages as the in situ synthesized glass slide-based arrays, as well as high feature redundancy, and have rapidly developed as a viable alternative for gene expression studies.

Because of the unique aspects of each of these different microarray platforms the nature of the primary data and how it should be processed to estimate gene expression levels varies between platforms. In this chapter, we focus on the preprocessing of data

derived from the Affymetrix GeneChip™ platform in some detail, since this platform is the most highly represented in the data repositories described above, with some discussion of how preprocessing approach should be modified for other technology platforms addressed briefly at the end.

A major weakness of all of the current gene expression microarray technologies is high variation in data quality. Systematic uncontrollable errors unrelated to target gene expression exist due to the nature of the experiments and are reflected in the raw data, which has a significant impact on subsequent statistical interpretations. Variability contributing to these errors may occur during the array production, target preparation, hybridization, array washing, or image collection stages of the experiment. The types of challenges that must be addressed include background noise, chip-to-chip biases and order-dependent patterns within an array. Biological variation within a subject (or between subjects) is often of interest as well and, when addressed properly by producing several replicate arrays per sample, introduces a need for additional processing. Limiting the presence and impact of such issues that can be quantitatively corrected after data collection is imperative to obtain accurate and reliable results from a microarray experiment.

Improper annotation of probes also contributes to unreliable data, such as the inaccuracy of probe set definitions for Affymetrix GeneChips due to utilization of out of date genome annotation (Dai et al. 2005). Another serious challenge with microarray experiments is systematic disturbances in chip design or manufacturing such as the order-dependent pattern discovered in Affymetrix GeneChips by Bjork and Kafadar (2007). The authors present indexed plots of typical values of interest such as expression values and variance that show a pronounced pattern consistent across several experiments. The study concludes that transcript order may delimit expression values and variance and thus results obtained from microarray data on specific Affymetrix platforms should be interpreted with caution.

Biological variability, although independent of the microarray experimental process, should also be considered when examining challenges and sources of error. The highest level of variability in the experiment is that between the subjects from whom the sample is obtained, but is generally of interest and should be estimated rather than eliminated. However, the variation within an individual should be accounted for by assaying replicate chips per subject and adjusting the analysis to account for such differences. Challenges that occur in the laboratory phase may seriously flaw the data and must be minimized prior to any analysis in order to avoid reporting inaccurate conclusions

B. Preprocessing of Primary Data

Controlling for the aforementioned challenges in a microarray experiment is imperative for production of reliable results and is typically addressed by preprocessing the probe-level data. Preprocessing occurs after feature extraction (image analysis) and alters the data in such a way that the resulting value is directly related to the expression level of a particular gene. Affymetrix chip preprocessing steps include probe annotation, background correction, normalization and summarization. Probe annotation involves the matching of array probes with mRNA transcripts and should also include the updating of transcript definitions based on the most current genome sequences. Background

correction is the process of using probe intensities to estimate the amount of background noise present on an array and adjusting for it. Normalization involves detecting and correcting for systematic non-biological differences between microarrays. Summarization combines probe intensities within a probeset into a single expression value.

High variation in data quality must be accounted for in order to arrive at accurate and consistent conclusions. Preprocessing methods focus on variations in experimental conditions rather than natural biological variations. The examination and subsequent adjustment of these differences play a crucial role in the overall goal of realizing the true difference in gene expression between subjects. The experimenter must be able to extract meaningful data characteristics from the noisy and improperly annotated primary data for use in downstream analysis such as differential expression, clustering and classification. For example, investigators rely on properly preprocessed data when predicting disease type. A patient may face serious consequences if given medication for the wrong disease type due to misclassification caused by an error in the background correction step of the microarray process. If the preprocessing methods are not reliable, then all subsequent analyses and conclusions may be flawed.

II. Current gene expression microarray preprocessing steps and methods – Affymetrix

A. Probe Annotation

The initial design of the probes included in many of the original Affymetrix microarray chips was based on incomplete genome sequences. For example, when the Affymetrix HG-U133 chip set was designed, the human genome sequence was only about 25% complete. In addition, even for more recently designed chips, as much as 30% of the PM probes may be problematic due to potential cross-hybridization and mis-annotation in the Affymetrix HGU-133A platform (Dai 2005, Harbig 2005, Shi 2006). The same annotation problem also exists with other microarray platforms (Shi et al., 2006).

Currently, even though the sequence of the human genome is complete, the manually curated reference sequences are being updated on a regular basis. To make sure that the probe annotation reflects the most current genome build, the Molecular and Behavioral Neuroscience Institute at the University of Michigan (BRAINARRAY, http://brainarray.mbni.med.umich.edu/Brainarray/) has established a process for the generation of periodically revised annotation files for all the Affymetrix platforms including human, mouse, rat, yeast, drosophila and Arabidopsis thaliana. The new annotations remove problematic probes such as those that cross hybridize with more than one transcript and those that do not hybridize to the correct target (Dai et al., 2005).

We have used Gene Ontology (GO) term co-clustering as an evaluation approach to assess the impact of using revised annotation on Affymetrix gene expression microarray data analysis (Kong et al., 2007). The evaluation approach is based on the premise that genes encoding proteins involved in the same biological process or protein complex will be coordinately expressed; that is, genes that have the same GO annotations are more likely to be clustered together (Lee et al., 2007). We have used the BRAINARRAY revised annotation and the Affymetrix original annotation to analyze several real biological data sets. Our results demonstrate that using revised annotation,

the p-values for the most significant GO terms in each gene cluster are much lower (i.e. more significant). In addition, the whole distribution of all the co-clustering p-values for all the GO terms is substantially lower when the revised annotation is used (see Figure 1). Thus, using revised annotation indeed produces much more significant co-clustering of related genes. Our result also supports the general approach of using GO term co-clustering as a useful evaluation metric to access preprocessing approaches with real biological data. We have applied the same methodology to assess the best microarray analysis pipeline (see detailed discussion below).

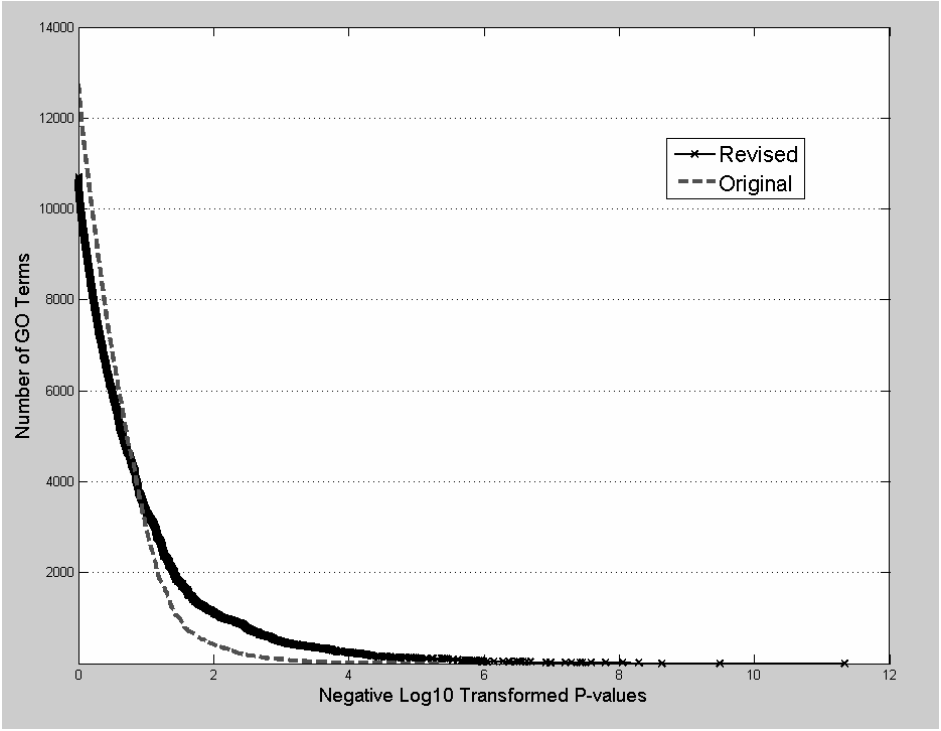

Figure 1. Comparison of GO term p-value distributions between original and revised Affymetrix probe annotation. The curve represents the distribution of all the p-values for all the GO terms in every cluster. The curve marked with "x" represents the co-clustering p-values using the revised annotation. The dashed line represents the co-clustering p-values using the original annotation.

B. Background Correction

It is known that the probe material that is affixed to the array carries some low level of background fluorescence. The same is true of the solid material to which the probes are attached. Therefore, there will be some fluorescence that is due to the material itself rather than the binding of the targets to the fixed probes.

It is also known that some probes bind imperfectly to their respective targets. There are two types of imperfect binding. Non-specific hybridization (NSH) is a type of background noise that is present when RNA fragments with sequences not meant for targeted probes bind to the probes anyway. The gene related to these probes will falsely appear expressed in the sample because the RNA bound to the unintended probes adds to the true signal. Cross-hybridization (XH) is the binding of a probe sequence to a target

that is at least partially identical to the true RNA target sequence of interest. Its effect can be partially attenuated by continual revision of annotation (see the previous section) and through the establishment of appropriately stringent hybridization conditions. But even under ideal conditions both NSH and XH cause spurious signal to be attributed to a probe, therefore falsely increasing its signal intensity.

The mismatch (MM) probes included in the Affymetrix chip design were originally thought to be a measure of background noise and non-specific hybridization. Since the MM intensities differ from the perfect match (PM) intensities in the middle base, a target that bound perfectly to the PM probe would, in theory, not bind to the MM probe. Therefore, one could simply subtract the MM value from the corresponding PM value in order to remove background signal and obtain a measure of specific hybridization (or true signal) for a particular probe.

The PM – MM estimation model for specific hybridization assumes two things, neither of which are true. First, the model assumes that the hybridization conditions are perfectly balanced so that a one-base mismatch is sufficient to eliminate hybridization for all probe pairs. Attaining such perfection in the conditions is impossible in practice. Second, it assumes that all MM intensities will be less than the corresponding PM intensities even if some NSH were to occur. However, it has been observed that approximately 30% of MM intensities are greater than PM intensities (Li and Wong, 2001a, Bolstad, 2004, Bolstad, et. al, 2003). The large MM intensities could be a result of cross-hybridization. Since we do not know the sequence for all genes in the human genome, it is possible that some of the MM probes are actually PMs for another gene. The MM intensities could also include effects of non-specific hybridization of the MM probe to the correct target for the PM, since the structure of the MM probes are so close to that of their PM partners. If that is the case, then there should be a correlation between the intensities of PM and MM probes.

Figure 2 shows two plots of different probe sets from the Affymetrix HG-U133A Latin Square Spike-In experiment (see http://www.affymetrix.com/support/technical/sample_data/datasets.affx for details). Four CEL files from the experiment were selected at random, and probe sets "203508_at" and "208010_s_at" were plotted for all four experiments. The top set of axes shows the intensity levels for probe set 203508_at, a spiked-in probe set. A probe set that is spike-in is one for whom the target concentrations were deliberately manipulated to achieve a known intensity for each experiment. The solid lines represent the logged PM intensities, and the dashed lines in the same color are the corresponding MM intensities. Note that for some experiments, the MM intensities are equal to or greater than the PM intensities. Furthermore, the MM intensities tend to track the PM intensities. In other words, there is evidence of correlation in the signals attributed to them most likely due to NSH.

The bottom panel of Figure 1 shows the PM and MM intensity values for probe 208010_s_at from the same four experiments. The probe 208010_s_at is not noted as a spike-in probe in Affymetrix's documentation for the experiment. However, probes 3 – 5 and 6 – 10 exhibit the behavior of a spiked-in probe in that they have much higher concentrations than would be expected, most likely due to XH. Furthermore, the MM intensities for these probes track the PM intensities, most likely due to NSH of the cross-hybridizing target. Thus, not only does this panel depict the correlation between PM and MM intensities, but also it shows evidence of cross-hybridization for several probes

within the probe set. It is reasonable to assume that other probes in different probe sets exhibit similar behavior.

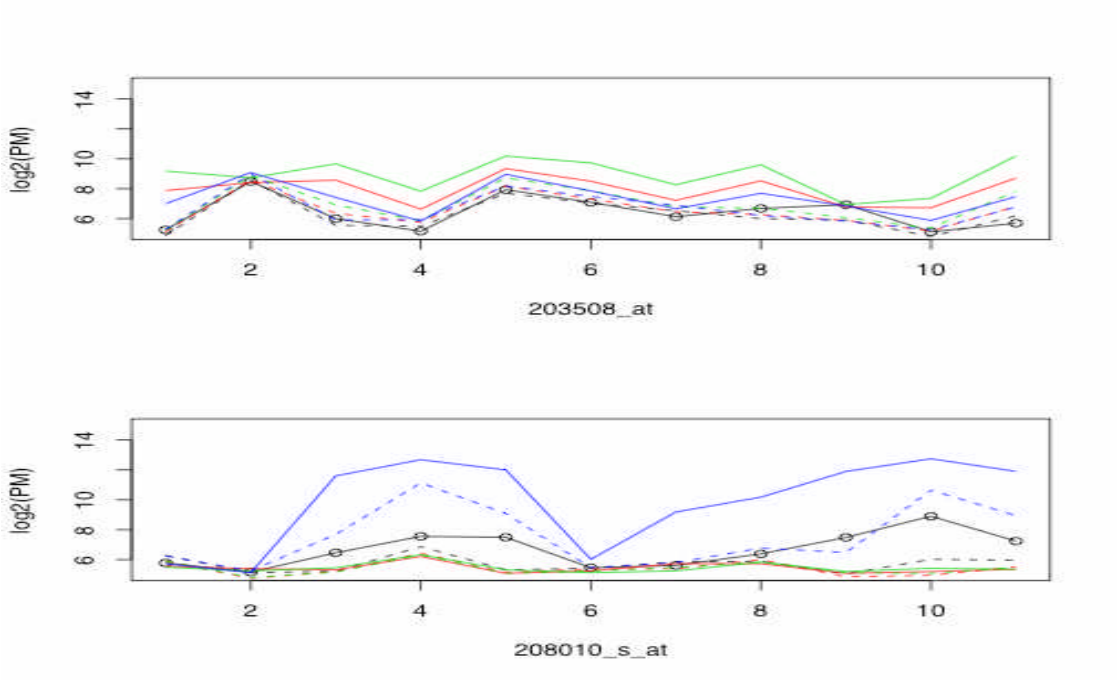

Figure 2. Log base 2 intensities for probe sets "203508_at" (top) and "208010_s_at"(bottom) from the HG-U133A Latin Square Spike-In Experiment. 203508_at is a spike-in probe, and 208010_s_at is not. The solid lines are the log bas 2 PM intensities and the dashed lines in the same color are the corresponding MM intensities. Note that the MM intensities tend to track the PM intensities very closely, indicating correlation between PM and MM intensities. Furthermore, several probes in the non-spiked in probe register high intensity, which is evidence of cross-hybridization.

Affymetrix attempted to correct the problem of large MM values with the "Ideal Mismatch" algorithm. This algorithm reduced MM values that are larger than their corresponding PM values by a given percentage, creating a MM value that is smaller than the PM partner. However, from Figure 2, it is clear that the MM probe intensities are correlated with the PM intensities. If a PM probe has high intensity, its corresponding MM does also, with very high probability. Therefore, subtracting their values from the PM values means that one is subtracting true signal from the PM probe intensity, giving artificially small estimates of the PM intensities.

Commonly used methods

MAS 5.0: MAS 5.0 (also called GCOS) is the software developed by Affymetrix to preprocess microarray data (Affymetrix 2001, 2003). As such, it is quite widely used, despite its poor performance compared to other methods (Li and Wong, 2001, Bolstad, et. al, 2003, McGee and Chen, 2006). The background correction algorithm for MAS 5.0 is often termed local adjustment or zonal adjustment. Each array is divided into k rectangular zones (the default for k is 16). The algorithm ranks the intensities for each pixel within each zone. The mean of the smallest 2% of the intensities is chosen as the background estimate for that zone. A standard deviation of the background is also

calculated. Both the background estimate and its standard deviation are smoothed, based on their distance from the center of each zone, to obtain one background estimate for the entire chip. The estimate is subtracted from each intensity measure at each point on the chip. A small fraction is added to any negative values that result from this process. The Ideal Mismatch algorithm is applied during the summarization step, to the background corrected data.

To its credit, MAS 5.0 attempts to make corrections for NSH. One might question the method, but at least the problem of NSH is acknowledged. The zonal background adjustment algorithm makes sense if one expects spatial correlation of background intensity within each chip. However, in the extreme situation of blotches and smears, the zonal adjustment algorithm will overcorrect in some places and undercorrect in others. If no spatial correlation is present, the zonal adjustment algorithm runs the risk of creating it.

One can also debate the wisdom of correcting each array within an experiment separately rather than as one unit. Other algorithms (e.g. RMA and DFCM) correct all arrays at once so that intensities from the correct arrays will be on the same scale, and therefore be comparable. However, one can imagine situations, such as when arrays for the same experiment are processed on different scanners or on different days, when it would be desirable to correct each array separately.

The main complaint about MAS 5.0 is the assumption that the PM and MM probes for a given probe pair have the same error. This assumption results in poor estimation of expression values, particularly at low RNA concentrations (Irizarry, Hobbs, et. al, 2003, Irizarry, Bolstad, et. al, 2003). The zonal adjustment algorithm also contains several arbitrary thresholds that must be determined, such as the number of zones, the smoothing parameter across the zones, and correction value for the Ideal Mismatch. Almost everyone who uses MAS 5.0 uses the default values, but there has not been a large-scale study on the effect of changing the default values for a particular set of data.

Model Based Expression Index (MBEI) The first challenger to MAS 5.0 was the Model-Based Expression Index (MBEI), also known as dChip (Li and Wong, 2001). It was shown to be more sensitive and specific than MAS 5.0 (Li and Wong, 2001, Bolstad 2003). Furthermore, the algorithm can detect defects in the chip and smooth them so that these aberrant intensities do not affect overall intensity estimates. The MBEI model is given by

$$y_{ij} = \mathrm{PM}_{ij} - \mathrm{MM}_{ij} = \theta_i \varphi_j + \varepsilon_{ij}, \text{where } \sum \varphi_j^2 = J.$$

Here, y_{ij} is the expression value of the j^{th} probe on the ith array, θ_i is the expression index for array i, φ_j is the probe-sensitivity index for probe j, and ε_{ij} is a random error term. The value of the probe sensitivity index determines whether or not the probe can be considered as aberrant, and therefore its intensity can be either down-weighted or excluded from calculation of the summarized gene intensity. MBEI does not estimate background directly, but rather computes an expression index for each probe. These expression indices are later normalized and summarized (excluding the "sensitive" probes) into a single expression value for each probe set.

Unfortunately, proper estimation of the parameters θ and φ requires at least ten arrays (Li and Wong, 2001a, b). Most microarray experiments consist of much smaller numbers of arrays. The MBEI algorithm produces unstable estimates for such

experiments. The normalization and summarization algorithms associated with MBEI are discussed in the next section.

Robust Multichip Average (RMA): The current most popular preprocessing algorithm for Affymetrix microarray data is Robust Multichip Average (RMA, Irizarry, et al, 2003). RMA, and many of its successors, are PM only algorithms, meaning that they utilize only the PM intensities from an experiment. MBEI can also be used with PM intensities only. The use of PM only was in response to the fact that the MM probes do not perform as theory predicted they would.

The background correction method used in RMA involves a convolution of an exponentially distributed signal and normally distributed noise. Specifically, it is assumed that X = S + Y, where X represents the observed PM intensities, S = the true signal that is exponentially distributed with rate parameter α, and Y = normally distributed noise, with parameters μ and σ. In reality, the distribution for the noise is a truncated normal. This is a normal distribution truncated at 0 so that there are no negative intensities. The point of background correction in RMA is to estimate the parameters μ, σ, and α, conditional on the knowledge of the observed intensities alone. The background correction is given by the following formula

$$E(S \mid X) = a + \sigma \left[\frac{\phi\left(\frac{a}{\sigma}\right) - \phi\left(\frac{x-a}{\sigma}\right)}{\Phi\left(\frac{a}{\sigma}\right) + \Phi\left(\frac{x-a}{\sigma}\right) - 1} \right]$$

where $\alpha = x - \mu - \sigma^2\alpha$, Φ represents the cumulative distribution function of the Normal distribution, $\phi()$ is the density function of the normal distribution (Bolstad, 2004). GCRMA (Wu, et. al, 2004) is a variation of RMA in which the background estimation attempts to account for the GC content of each probe. Therefore, the background correction is not based on the exponential-normal model. The normalization and summarization algorithms for GCRMA are the same as they are for RMA. In general, GCRMA does not perform well (Chen, et. al, 2007), and it will not be discussed further.

There are two main criticisms of the exponential-normal model. First, the parameters are estimated in a very ad hoc manner. Second, the exponential-normal assumption is impossible to check in practice, since we cannot separate noise from signal for real data. Furthermore, even if the data follow an exponential-normal convolution model, the procedure devised to estimate the parameters results is extremely poor estimates (McGee and Chen, 2006).

RMA-Mean and RMA-75: Since the estimation procedure used for RMA in the Bioconductor (Gentleman, 2004) package produces poor estimates, McGee and Chen (2006) devised different methods to estimate the mean (μ) and standard deviation (σ) of the background noise while still assuming an exponential-normal convolution model. The best performing method was one in which a one-step iterative formula was employed using the following mathematical relationship between μ, σ. α, and the mode, x_m.

$$\phi\left(\frac{x_m-\mu}{\sigma}-\alpha\sigma\right)=\alpha\sigma\left[\Phi\left(\frac{x_m-\mu}{\sigma}-\alpha\sigma\right)\right]$$

The initial estimates for the iteration were given by the estimates from the ad hoc estimation algorithm.

Recall that RMA uses the mode of the data greater than x_m to estimate the rate parameter of the exponential distribution (α). McGee and Chen examined the following statistics calculated from the intensities greater than x_m as replacements for the RMA estimate: the mean, the median, the seventy-fifth percentile, and the 99.95th percentile. For all estimates, the mean-squared error (MSE) between the estimated value and the known value was calculated. The MSE measures precision (variability) and accuracy (bias) simultaneously.

Table 1 shows the results in which an exponential-normal convolution model with parameters $\mu = 30$, $\sigma = 5$, and $\alpha = 5$ was generated. An examination of several data sets indicates these parameter values are reasonable for a variety of microarray data. The column headings in the table refer to the different methods of estimating α. Estimation for μ and σ was done via the one-step iteration method described previously. Other simulations using different values of the true parameters were done with the same qualitative results.

	Estimation Method				
True	Ad Hoc	Mean	Median	75th %-tile	99.95th %-tile
Parameter	233	2.62	2.62	2.62	2.68
Values	92	1.35	1.35	1.51	1.39
$\mu = 30$	45135	4.70	10.2	2.56	22.1
$\sigma = 5$					
$\alpha = 5$					

Table 1. Mean Squared Errors for parameter estimates via the RMA ad hoc method and four other methods for estimating μ, σ, and α. Smaller values for the MSE indicate better estimates. The ad hoc method (first column) gives very poor estimates. All other methods perform well in comparison. The best method is the one using the 75^{th} percentile of all intensity values greater than the overall mode of all intensities to estimate α. The method where the mean of these values is used is a close second.

Since the one-step iterative estimate using either the 75^{th} percentile or the mean of the data greater than the overall mode of all intensities to estimate α performed best, McGee and Chen devised two new background correction methods, based on the exponential normal model, but using different estimates for the parameters. The methods are called RMA-Mean and RMA-75. For both methods, the mean and standard deviation of the noise distribution are estimated using the one-step iterative formula given previously. The rate parameter for the exponential distribution is estimated with the mean of all intensities greater than x_m (RMA-Mean) or the 75^{th} percentile of all intensities greater than xm (RMA-75).

RMA-Mean and RMA-75 were tested using the Affymetrix Spike-In data to ascertain whether better estimates of the parameters produced better results in terms of sensitivity and specificity in the selection of differentially expressed genes. Indeed, these

two methods outperformed the original RMA method (McGee and Chen, 2006), although not by as much as expected given the clear superiority of the estimators used.

Distribution Free Convolution Model (DFCM): Estimates that have dramatically smaller MSEs than the original ones should produce great improvements in sensitivity and specificity *if the underlying model is correct*. Using the Affymetrix Latin Square spike-in data sets, McGee, et. al, (2006) showed that the exponential distribution was too light-tailed to model the signal adequately, and that the distribution of the noise was likely non-normal, also. Therefore, the exponential-normal model is not a good fit for gene expression microarray data.

The idea of a convolution model for gene expression is reasonable, even if the distributional assumptions attached to it are not. A convolution suggests that noise pervades the signal at all levels, which is likely the case with gene expression data from microarrays. Therefore, McGee et. al (2006) suggested a convolution model in which signal and noise parts of the convolution were estimated using nonparametric (or distribution free) methods. The method is called the Distribution Free Convolution Model (DFCM). It has the advantage that no underlying distributions of the signal and noise are assumed. It also makes some use of the MM intensities, which have been ignored in most preprocessing methods up to this point.

The DFCM procedure for background correction proceeds as follows:

1. Obtain the smallest q_1 percent of the observed PM intensities (X). ($q_1 < 30\%$ works best).
2. Obtain the smallest q_2 percent (typically 90% or 95%) of MM intensities associated with the PM intensities gathered in step 1. These MM intensities are assumed to measure background noise.
3. Use a nonparametric density estimate of the lowest q_2 percent of the MM intensities to find the mode of the noise distribution. The mode is used to estimate the mean of the noise distribution.
4. Estimate the standard deviation of the background noise by calculating the sample standard deviation of MM intensity values that are smaller than the estimated mean.
5. Correct the intensity for each probe in each probe set by subtracting the estimated mean. For very small intensities, a scaled value of the mean is subtracted.

The DFCM procedure avoids the erroneous distributional assumptions of RMA. Furthermore, it has been shown to produce good results in terms of sensitivity and specificity, in combination with various normalization and summarization methods. These results hold for real data and for spike-in data.

C. Normalization

Normalization is required in order to correct chip-specific biases influencing all probes on an array. Such biases include, but are not limited to, printing, hybridization, or scanning artifacts. In this section, we describe commonly used methods, their advantages, and their disadvantages. Schuchhardt, et. al (2000) and Hartemink, et. al (2001) give more complete lists of sources of obscuring variation.

Housekeeping Genes: So-called "housekeeping genes" have also been used as a basis for normalizing microarray data. Affymetrix, as well as other manufacturers, include several such genes on each array. These probe sets are not supposed to catch signal from the target genes; therefore their intensities should be the same (near 0) across all arrays. However, it has been shown that housekeeping genes vary more than expected (Quackenbush, 2002). Furthermore, there are generally not enough of them to determine a reasonable function for adjusting the other probes.

Constant Normalization: The earliest method of normalization was called constant or global normalization. The principle of this type of normalization is to multiply the intensity values for all arrays in an experiment by a constant so that the mean (or median) for all arrays is the same. There are many variations on this theme. For example, one can find a constant such that the resulting multiplication gives a mean of 1 for all arrays (and thus a log base 2 mean equal to 0). This is done by dividing each value in an array by the overall mean of the array. Another method is to choose a baseline array, find its mean, and then multiply all other arrays by a constant so that they have the same mean as the baseline array. This is the normalization scheme for MAS 5.0.

Constant normalization has been shown to perform poorly compared to other normalization methods (Yang, et. al, 2002). The variation of the procedure used for MAS 5.0 has been criticized because of the necessity of choosing a baseline array, and this choice is arbitrary for most experiments (Bolstad, et. al, 2003). Furthermore, constant normalization is a linear algorithm. Therefore, it operates under the assumption that non-biological variation adds linearly to the true intensity levels. Given the complex nature of microarray experiments, this is probably not the case.

Invariant Set: To address the premise that extra-biological effects are non-linear in nature, invariant set normalization (Li and Wong, 2001a, b) uses non-linear regression to map values from each array in an experiment to a baseline array. The baseline array is defined as the array having "median overall brightness". Once a baseline array has been selected, normalization proceeds by finding a set of genes whose intensities do not vary in rank order from the baseline array to the comparison array (rank invariant set). Cross-validated smoothing splines are used to define a non-linear relationship between the invariant set and the comparison array, then all probes on the comparison array are adjusted using this fit. Details of the fit and the selection of the invariant set of genes are given in Li and Wong (2001b). Invariant set normalization suffers from the same problems as do other baseline methods. In addition, it tends to be computationally expensive, as one must fit (and cross-validate) a complicated non-linear model to each array in an experiment.

Contrast: Another non-linear method is contrast normalization (Åstrand 2003). It is based on the M versus A plot, where M is the difference in log expression values and A is the average of log expression values (Dudoit, et. al, 2002). The data are placed on a log scale and the basis is transformed. After transformation, a series of n -1 (where n is the number of arrays) curves are fit to the M versus A curves for a pair of arrays, with the goal of adjusting the intensities for the two arrays so they are centered on the line where $M = 0$. The normalized data is obtained by back transforming to the original basis and

exponentiating (Bolstad, et. al, 2003). Other than its complexity, the main disadvantage of the contrast method is that it is computationally expensive since the algorithm is performed on the set of pairwise combinations of arrays.

Local Linear Smoothing Splines (LOESS): LOESS is more computationally expensive than the contrast method, but it seems to have enjoyed more popularity. It was introduced in the context of cDNA microarray data (Dudoit, et. al, 2002), and is also based on the M versus A plot. For LOESS, a localized linear regression function is fit to the M versus A plots for all pairwise combinations of arrays. The linear regression is used to adjust the intensities so that the data fall around a horizontal line (where M = 0). Usually, the normalization is performed using rank invariant sets of probes in order to decrease run time (Bolstad, et. al, 2003). Once the program has computed normalized values for each pair of arrays, these values are summarized and applied to the set of all arrays. Up to 2 iterations of the process are done in order to assure that further adjustments are minimal.

Variance Stabilizing Normalization (VSN): VSN (Huber, et. al., 2002) was devised to account for the fact that the variance of microarray data is dependent on the mean. Specifically, variability of microarray data increases as the intensity increases. For large intensities, the relationship between the mean and standard deviation is nearly linear. However, this is not the case for smaller intensities, often due to the fact that intensities below a certain arbitrary threshold are dismissed as probes that are not expressed ("absent" in the parlance of Affymetrix analysis). VSN uses a model that assumes both multiplicative and additive error. The data are transformed so that the variance of the intensities is equal regardless of the mean. The transformation is a hyperbolic sine function, which has a relatively simple mathematical relationship to the logarithm. VSN is grounded in good mathematical theory, yet it has not enjoyed the popularity of some of the more ad hoc methods, mainly because it has been shown to underperform on spike-in data (Bolstad, 2004). It may also be the complexity of the theory that keeps practitioners from readily employing this method.

Quantile Normalization: Quantile normalization is a multi-chip method in which all arrays are made to have the exact same distribution. The method proceeds as follows

1. Sort the intensities for all arrays in the experiment. Find the maximum intensity in each array.
2. Substitute the median of all maximum intensities for the maximum intensity on each array.
3. Obtain the next-to-largest intensity for each array. Calculate its median. Substitute the median for each probe having that next-to-largest intensity.
4. Continue until the smallest intensity is reached.

Figure 3 shows a simple example of four arrays with four probes each to illustrate the method. The set of brackets to the far right represents the original data. The rows are probe intensities, and the columns represent arrays. The maximum value for each array is 9, 8, 6, 9. The median of these values is 8.5. In the second set of brackets, each of the previous maximum values have been replaced by their median: 8.5. The second-largest values are 8, 7, 6, and 8. In the third set of arrays, these values are replaced by their

median: 7.5. The process continues until all of the original intensity values have been replaced.

$$\begin{bmatrix} 5 & 8 & 4 & 8 \\ 3 & 2 & 7 & 9 \\ 8 & 3 & 2 & 7 \\ 9 & 7 & 6 & 2 \end{bmatrix} \Rightarrow \begin{bmatrix} 5 & 8.5 & 4 & 8 \\ 3 & 2 & 8.5 & 8.5 \\ 8 & 3 & 2 & 7 \\ 8.5 & 7 & 6 & 2 \end{bmatrix} \Rightarrow \begin{bmatrix} 5 & 8.5 & 4 & 7.5 \\ 3 & 2 & 8.5 & 8.5 \\ 7.5 & 3 & 2 & 7 \\ 8.5 & 7.5 & 7.5 & 2 \end{bmatrix} \Rightarrow \begin{bmatrix} 4.5 & 8.5 & 4.5 & 7.5 \\ 2.5 & 2.5 & 8.5 & 8.5 \\ 7.5 & 4.5 & 2.5 & 4.5 \\ 8.5 & 7.5 & 7.5 & 2.5 \end{bmatrix}$$

Figure 3: A simple example of quantile normalization. The original data set with four arrays (columns) and four probes on each array (rows) is given on the far left. The algorithm begins by taking the maximum values for each array (9, 8, 7, 9, respectively). It calculates their median (8.5) and then replaces each maximum value with the median (second array). The third array shows the replacement of the second largest intensity in each array by their median (7.5). The array on the far right is the finished product.

Note that quantile normalization not only results in arrays with the same distribution, but also the arrays have the exact same values (in different places). It is possible that quantile normalization will overcorrect for non-biological bias. There is some evidence that it greatly reduces the intensity values of probes with high intensities, and increases intensity values of probes with low ones (Calza, et. al, 2008), thus making differentially expressed genes more difficult to detect.

Evaluation of Methods and Disclaimers: Of the methods mentioned above, LOESS performs the best in terms of sensitivity and specificity, according to GO term co-clustering (Kong, et. al, 2007). Its computational complexity is not as much of a burden as it was even six years ago. However, Calza, et. al (2008) note that LOESS suffers from the same problem as quantile normalization in that it tends to shrink high and low intensities toward the mean of all intensities. In order to understand the tendency of LOESS and quantile normalization to shrink estimates, it is important to understand the assumptions underlying all current normalization methods.

There are scenarios in which one might expect a large majority of the genes to vary in intensity between different arrays. However, all of the normalization algorithms described in this section assume that less than 10-20% of genes vary in intensity between arrays. In fact, all normalization algorithms must have some set of observations that are relatively stable. Otherwise, there is no standard to which to perform the normalization. The normalization methods listed above use all intensities on the arrays to obtain normalized values. This is fine as long as not many of the genes change between arrays.

The second assumption is that there are equal numbers of up and down regulated genes in the experiment (symmetric differential expression) (Bolstad, et. al, 2003). There are experiments in which we expect most of the genes to be up-regulated (or down-regulated). If this assumption is violated, it no longer makes sense to center normalized values around a horizontal line. Therefore, in experiments where there are a large majority of up (or down) regulated genes, attempting to force their intensities to center around a horizontal line would result in grievous loss of signal, and thus a loss of power to detect differentially expressed genes.

Unfortunately, these assumptions are rarely checked in practice. In fact, it would be extremely difficult to do so. If either of these assumptions is violated, the normalization algorithms described in this section may introduce more bias than they correct. It has also been shown that choice of normalization algorithm has a great influence on the outcome of the experiment in terms of which genes are declared as differentially expressed (Hofmann, et. al, 2002). Therefore, when choosing a normalization algorithm, one must carefully consider whether the aforementioned assumptions can be expected to hold. A recent algorithm, called least variant set normalization (Calza, et.al, 2008), claims to produce more stable results when these assumptions are violated, at least for Spike-In data. Since it is a new algorithm, it has not been tested on large numbers of data sets as the other algorithms have.

D. Summarization

Summarization methods are necessary for Affymetrix arrays in order to establish a single expression value corresponding to a particular gene. Some of the more common methods are average difference (Affymetrix 2001, Tukey bi-weight (Affymetrix 2001, 2003, Li-Wong (Li and Wong, 2001 a,b) and median polish (Bolstad, 2004).

Average difference (AvgDiff) was the method of summarization devised by Affymetrix for MAS 4.0 (Affymetrix, 1996). It is defined as the sum of the differences between the PM probe intensities and their corresponding MM intensities for a probe set, divided by the number of "present" probes in that set. The original GeneChip software included all pairs. MAS 4.0 excluded outlier pairs of PM – MM more than three standard deviations from the mean PM – MM value. AvgDiff can be negative if MM > PM. Furthermore, it is in general unwise to subtract MM values from PM values, since the MM values tend to carry some signal (see Figure 1). AvgDiff is not log-transformed; therefore, the variance of the summarized values tends to increase with the mean. Affymetrix recommends the use of MAS 5.0 over MAS 4.0 (Affymetrix 2001). In general, AvgDiff is no longer used for summarization.

MAS 5.0 employs a robust method, based on the *Tukey-Biweight* function, to summarize each probe set. First, the probe intensities are background corrected, but not normalized. In MAS 5.0, normalization comes after summarization. Then an ideal mismatch is calculated if MM > PM. The ideal mismatch is the PM intensity minus some correction value, which depends on the magnitude of the difference between the MM and PM intensities for a single probe pair. Next, the adjusted PM – MM intensities are log-transformed, and the Tukey-Biweight function applied to obtain the summarized expression value. The Tukey-Biweight function is a weighted calculation, meaning that values that are farther from the median log(PM – IM) intensity receive lower weight than those close to the median. Therefore, the influence of outliers is minimized, both by the use of the median as a measure of center and by the weighting scheme. However, this method has also fallen out of favor, mainly because of the use of the ideal mismatch.

PLIER is the latest algorithm developed by Affymetrix (Affymetrix 2005). The ideal model for the expected value of the observed binding for PM and MM probes is given by:

$$E(PM_{ij}) = \mu_{ij} = a_i c_j + B_{ij}$$
$$E(MM_{ij}) = B_{ij}$$

where B_{ij} is the background binding for probe pair i on array j, μ_{ij} is the binding level of probe i on array j, a_i is the binding affinity of probe i, and c_j is the concentration of RNA within sample j (Therneau and Ballman, 2005). In this formulation, the background for the PM intensities is equal to that of the MM intensities, and the MM intensities carry no signal. There are random errors associated with both PM and MM binding, ε_P and ε_M, respectively, and it is generally accepted that these errors are multiplicative (which is why log transforms are recommended). MAS 5.0 makes the assumption that $\varepsilon_P = \varepsilon_M$. The limitations of this assumption were discussed earlier.

PLIER assumes that $\varepsilon_P = 1/\varepsilon_M$, which is a biologically implausible assumption, certainly in the light of Figure 1, where it is clear that the PM and MM probe intensities have a positive correlation in general. Yet, PLIER outperforms MAS 5.0 for benchmark data (Cope, et. al, 2004). PLIER also obtains more accurate expression values for low RNA concentration levels. Therneau and Ballman (2005) examine the ability of PLIER to estimate the binding affinity for a single probe at low concentrations of RNA. The function produced by PLIER for estimating the ideal error curve (when all assumptions are true) has the correct shape, at least when considering a single probe. However, they speculate that differences in the observed error functions from the ideal error functions will be greatly increased when summarizing the intensities over an entire probe set. Therneau and Ballman (2005) go so far as to state that the improved performance of PLIER over MAS 5.0 is simply due to "good fortune". Further investigations of the performance of PLIER have shown that it is clearly outperformed by RMA, MBEI, DFCM, and many others (McGee, et. al, 2006).

The MBEI (dChip) algorithm computes summarized values using the following formula

$$\tilde{\theta}_i = \left(\sum_j y_{ij} \phi_j \right) \Big/ J,$$

where J = the number of probes in the probe set (excluding outliers), y_{ij} is the observed intensity level for probe j on array i, and ϕ_j is a measure of probe effect (Li and Wong, 2001b). θ_i and ϕ_j are determined using an iterative algorithm, where each value is assumed fixed while the other is estimated. The algorithm stops when changes in successive estimates of the parameters are sufficiently small. Even with outliers removed, the summarization scheme for MBEI gives the most weight to probes with the largest intensities because it is a sum of the observed intensities (i.e. it is not robust). Given the presence of NSH and XH in the data, intensity values are often overestimated. This could account for the poor performance of MBEI in comparison with more recent methods (Bolstad, et. al, 2003, McGee and Chen, 2006).

Factor Analysis for Robust Microarray Summarization (FARMS): Hochreiter, et. al, (2006) developed, FARMS, another summarization method in which the observed values are modeled with a linear model. Specifically, let x = the zero-mean normalized PM intensities for a given probe set, and z = RNA concentration. The observed values are assumed to depend on concentration via the relationship

$$x = \lambda z + \varepsilon, \quad z \sim N(0,1) \quad \text{and} \quad \varepsilon \sim N(0, \Psi)$$.

The error term is assumed to have a multivariate Normal distribution, and Ψ is a diagonal variance-covariance matrix. The terms z and ε are assumed to be statistically independent. FARMS performs better than other competitors using benchmark data (Cope, et. al, 2004), particularly where sensitivity and specificity are concerned. It is also computationally efficient, as the EM algorithm is used to estimate the model parameters. However, Chen, et. al, 2007 show that FARMS does not perform as well as RMA or DFW, although this result is counter to previous claims.

Median Polish: The summarization method for RMA is median polish (Tukey, 1977). It is a robust method for fitting the following linear model

$$\log_2\left(y_{ij}^{(n)}\right) = \mu^{(n)} + \theta_j^{(n)} + \alpha_i^{(n)} + \varepsilon_{ij}^{(n)}$$

with constraints median(θ_j) = median(α_i) = 0 and median$_i$(ε_{ij}) = median$_j$(ε_{ij}) = 0. Here, the superscript (n) represents the nth probe set on array j, y_{ij} refers to the observed intensity of the i^{th} probe, α_i represents a probe effect, and θ_j is an array effect (Bolstad, 2004). Median polish begins by arranging the data in a matrix for each probeset n such that the probes are in rows and the arrays are in columns. A column of zeros is appended to the right of the matrix, and a row of zeroes is appended to the bottom. Next, the median of each row (ignoring the last column of zeros) is calculated, subtracted from each observation in the row and added to the final column. The procedure proceeds similarly for the columns. The algorithm continues until the changes are arbitrarily small. The end result is estimates for μ, θ, α, and ε, which are used to compute a summarized value for each probe set.

For a large number of arrays, the median polish algorithm will be computationally expensive. Median polish does not provide standard error estimates in a natural way, and it can be applied only in balanced row-column effect models (Bolstad, 2004). Furthermore, median polish gives different answers depending on whether the procedure begins with rows or columns.

Distribution Free Weighted Summarization: With the exception of MBEI, none of the summarization methods mentioned above takes into account the fact that some probes perform poorly. They either tend to cross-hybridize with the target, resulting in too much signal, or they do not catch signal at all. Evidence of both kinds of probes is seen in Figure 1. MBEI tends to favor probes with large intensities, which means that expression values will be overestimated. Furthermore, the MBEI model assumes independent and identically normally distributed error terms for ease of estimation. For real data, there may be multiplicative error, or the terms may be correlated.

Chen, et. al (2007) devised a method that eliminates both cross-hybridizing and under-performing probes in the summarization process. The method, Distribution Free Weighted summarization (DFW) is data driven method that obtains an estimate of standard deviation for each probe across all arrays, regardless of the experimental manipulation done to the array. Probes that have large standard deviations tend to be either differentially expressed between treatments, cross-hybridizing, or underperforming. The algorithm for DFW uses separate measures of variability to determine to which group a probe belongs. For cross-hybridizing or under-performing

probes, a small weight is assigned during summarization so that the intensity levels assigned to these probes contribute little to the final expression value for that gene. The elimination of such probes has proved fruitful, as DFW was able to identify all differentially expressed genes in the Affymetrix Latin Square HG-U95Av2 spike-in experiment with no false positive (AUC = 1.0), and all DEGs from the HG-U133A experiment with only two false positives (Chen, et. al, 2007). Furthermore, DFW performs best on real data, as evidenced by the GO algorithm (Kong, et. al, 2008).

E. Present/absent calls

It is not likely that all genes on a chip are represented in a given target sample and those that are not should be filtered out of the preprocessed dataset before further analysis. Reducing the abundant number of features after background correction, summarization and normalization are carried out alleviates problems that may be encountered in large sample statistical analysis. A process that calls a gene present or absent based on its signal to noise ratio is commonly utilized to determine whether a gene is actually expressed in the sample. An absent call for a gene indicates there was not enough mRNA transcript in the sample to bind to the probes sequenced for that gene and the gene is filtered out. However, if a gene is called present then there was sufficient mRNA for the gene to be detected and the gene is kept for subsequent statistical analysis.

The current detection algorithm for Affymetrix arrays uses a discrimination score and its *p*-value for assigning present, marginal and absent calls. The discrimination score, R=(PM-MM)/(PM+MM), where PM is perfect match intensity and MM is mismatch intensity, is calculated for a probe pair and then compared to a user-definable threshold τ (default 0.015). The one-sided Wilcoxon's Signed Rank test is used to obtain a *p*-value (Affymetrix 2007). The user then defines cutoff values such that if a *p*-value is less than α_1 the gene is present, between α_1 and α_2 the gene is marginal and greater than α_2 the gene is absent, where the default values of α_1 and α_2 are 0.04 and 0.06, respectively. The experimenter may choose to use marginal genes in further analysis or consider them absent.

The Affymetrix Statistical Algorithms Reference Guide claims that with this method, called MAS5, the smaller the *p*-value the more likely the measured mRNA transcript is detected. However, a positive correlation often exists between the perfect and mismatched probe pairs such that a small *p*-value actually provides evidence of this relationship rather than evidence of detectable levels of expression as shown in Figure 4. Another issue with the MAS5 approach is that the set of genes called marginal actually includes genes that should be called present and genes that should be called absent (see Figure 5). If the marginal genes are chosen for use in downstream analysis, the experimenter is unnecessarily increasing the number of false positives. If the marginal genes are filtered out, then the number of false negatives will be falsely inflated.

The discrimination score, R, uses the mismatch intensities as a baseline for detecting signal in perfect match intensities. This is problematic due to the nature of the GeneChip experiments. The MM probe is meant to be a control that measures a corresponding PM's nonspecific hybridization and the PM probe sequence is intended to hybridize with the mRNA from a specific gene. However, the PM probe can experience nonspecific binding and the MM probe can hybridize with the mRNA meant for the PM probe. Thus the intensity measured is not simply a signal plus background noise, rather a

combination of the signal, unavoidable nonspecific binding and background noise. Simply subtracting PM-MM will not result in a properly corrected PM signal and therefore the *p*-value corresponding to the calculated R is not a reliable measure of detection.

A new method of calling presence or absence of a gene should utilize the fact that MM probes are not reliable estimates of background noise of PM probes because of non-specific hybridization to the MM probes themselves. Warren et al. address the issue with a method called "Presence-Absence calls in Negative Probesets" that utilizes Affymetrix-reported probes that should not bind to intended sample RNA sequences (Warren 2007). Although the authors claim the method performs better than the current Affymetrix detection algorithm, it was developed for only two Affymetrix platforms. A method that may be applied to any platform, including those without the MM feature of Affymetrix chips, was suggested by Wu and Irizarry but uses only PM probe information (Wu and Irizarry 2005). This *half-price* method makes detection calls without the use of MM probes and therefore may be less reliable because not all available information is used. Additional approaches to detecting gene expression in microarrays should be formulated and tested to provide microarray users with more accurate methods of filtering genes with low signal to noise ratio.

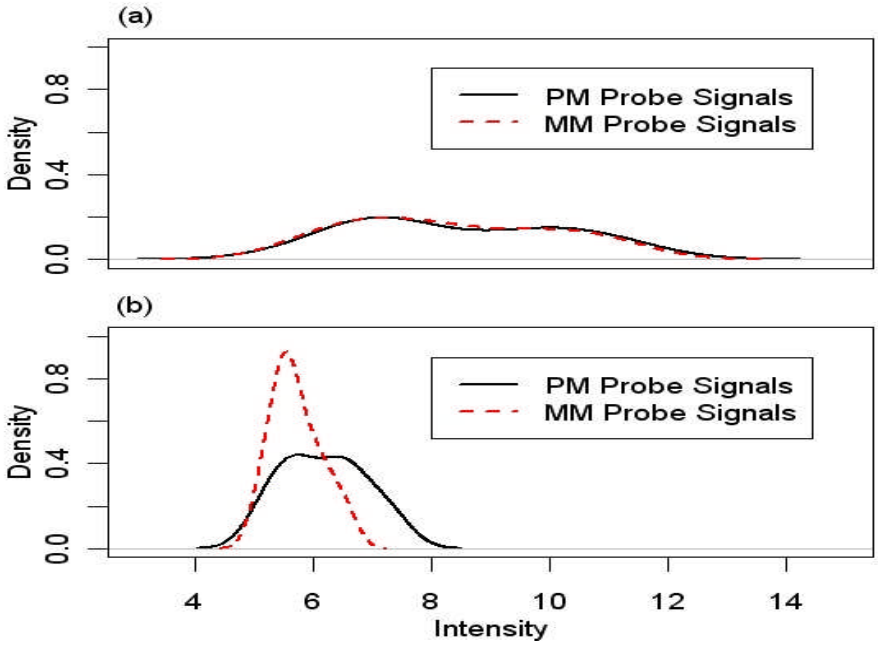

Figure 4. Intensity distributions for 'present' and 'absent' genes. (a) The densities of $\log_2$ PM and MM probe intensities for a single gene from a single study. Although there is signal present in this probeset, the Affymetrix detection algorithm (MAS5) calls this gene absent because the PM and MM distributions are so similar. (b) The densities of $\log_2$ PM and MM probe intensities for a single gene from a single study. The Affymetrix detection algorithm calls this gene present because the PM and MM intensity distributions are different.

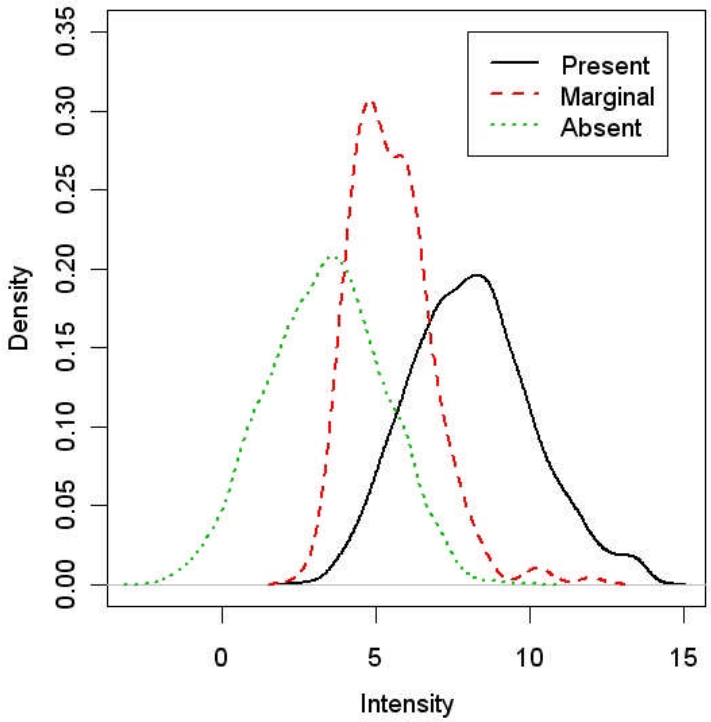

Figure 5. Distributions of $\log_2$ intensities for genes called present, absent and marginal genes from a single experiment. The two peaks in the marginal genes indicate there should further separation of present and absent genes. Note there are negative values after taking the logarithm of a measured signal that is less than 1.

III. Preprocessing steps and methods for other platforms

Unique aspects of each of the platforms require unique approaches to preprocessing. For example, the spotted arrays utilize several pins to spot transcripts onto slides, whereas the photolithography methods construct the oligonucleotides one nucleotide at a time directly on the array. A pin (or print-tip) bias may be present in spotted arrays, and must be accounted for, but is absent in *in situ* arrays. Also, the background estimation requires image analysis software due to the nature of the spotting. The spots are not uniform within or between spotted arrays, introducing issues with defining foreground and background. See Gentleman (2005) for more details on preprocessing methods for spotted cDNA and long oligonucleotide arrays. Microbead-based microarrays also require platform specific preprocessing steps which are discussed thoroughly by Dunning, et al. (2008).

IV. Preprocessing Methods Evaluation

Currently, there are many microarray analysis methods available which makes it challenging for scientists to decide which method to use for their data analysis. The quality and the credibility of these methods when applied need to be assessed fairly. To assess the performance of these algorithms, a series of data sets were produced in which a group of known transcripts were mixed at known quantities and their levels in the mixture measured using standard microarray methodologies (Irizarry et al., 2003). These

so-called "spike-in" data sets provide the ability to assess sensitivity and specificity by comparing the receiver operating characteristic (ROC) curves produced (McGee and Chen, 2006). While this approach is useful, there is some concern that analytical approaches that work well with spike-in data may not work as well with data derived from real, complex biological samples. ROC analysis is not feasible for real biological data since the true expression values for target mRNA's are rarely known, and so methods of comparison other than ROC curves are needed.

The Gene Ontology, one of the most successful biomedical ontologies, includes biological process, molecular function and cellular component terms linked together in a directed acyclic graph with "is_a" and "part_of" relationships. The GO has been used extensively to annotate prokaryotic and eukaryotic gene products based on information described in the scientific literature. Based on the premise that an improvement in any step of microarray data analysis should be reflected in improved co-clustering of related genes, we have successfully applied GO term co-clustering as a comparative tool and assessed the impact of using revised annotation on Affymetrix gene expression microarray data analysis using real biological data (Kong et al., 2007).

We have applied the same methodology to evaluate various background correction, normalization and summarization methods. We have chosen to assess several popular or newly developed methods including five background correction methods, five normalization methods and four summarization methods using real biological data. Interestingly, there exist some interdependencies among these methods. We have assembled about 300 pipelines which represent all possible combinations of background correction, normalization and summarization methods. Through evaluating their GO term co-clustering characteristic, we conclude that RMA, RMA-Mean, RMA-75 and DFCM are better background correction methods compared to MAS. For normalization, loess and invariant set outperform constant, contrasts and quantiles. Median polish is the best summarization method compared to DFW, FARMS and MAS5 (Kong et al., 2008). For the dataset that we tested, the best analysis pipeline is RMA-75 as background correction, invariant set as normalization and DFW as summarization method. Our future goal is to use a data-driven approach to provide the best analysis pipeline for different distributions of microarray data, which will meet the need of the scientific community.

References

Affymetrix (2003). Technical note: design and performance of the GeneChip Human Genome U133 plus 2.0 and Human Genome U133A plus 2.0 arrays.

Affymetrx (2005). Guide to Probe Logarithmic Intensity Error (PLIER) Estimation.

Affymetrix (2007). Statistical Algorithms Reference Guide, Data Analysis Fundamentals Technical Manual.

Åstrand M (2003). Contrast Normalization of oligonucleotide arrays. *Journal of Computational Biology* 10 (1): 95-102.

Bjork, K. and Kafadar, K. (2007). "Order dependence in expression values, variance, detection calls and differential expression in Affymetrix GeneChips," *Bioinformatics* 23(21): 2873-2880.

Bolstad BM. (2004). Low Level Analysis of High-density oligonucleotide array data: Background, normalization and summarization [dissertation]. University of California at Berkeley.

Bolstad BM, Irizarry RA, Astrand M, Speed TP. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19:185-193.

Calza S, Valentini D, and Pawitan Y (2008). Normalization of oligonucleotide arrays based on the least-variant set of genes. *BMC Bioinformatics* 9: 140.

Chen Z, McGee M, Liu Q, and Scheuermann RH (2007). A Distribution Free Summarization Method for Affymetrix GeneChip Arrays. *Bioinformatics* 23:3:321-327.

Choe S, Boutros M, Michelson A, Church G, Halfon M. (2005). Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biology*, 6(2):R16.

Cope LM, Irizarry RA, Jaffee H, Wu Z, and Speed TP (2004). A benchmark for Affymetrix GeneChip Expression Measures. *Bioinformatics* 1(1):1 – 13.

Dai, M., Wang, P., Boyd, A.D., Kostov, G., Athey, B., Jones, E.G., Bunney, W.E., Myers, R.M., Speed, T.P., Akil, H., Watson, S.J. and Meng, F. (2005). Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.* 33(20): e175.

Dunning, M., et al. (2008). Statistical issues in the analysis of Illumina data. *BMC Bioinformatics* Vol. 9, No. 1.

Dudoit S, Yang YH, Callow MJ, and Speed TP (2002). Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments. *Statistica Sinica* 12(1): 111 – 139.

Gentleman RC, Carey VJ, Bates DM, et al. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology* 5:R80.

Gentleman, RC, et. al (2005). Bioinformatics and Computational Biology Solutions Using R and Bioconductor. New York: Springer.

Harbig, J., Sprinkle, R., and Enkemann S.A. (2005). A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array. *Nucleic Acids Res.* 18;33(3):e31.

Hartemink A, Gifford D, Jaakkola T and Young R (2001). Maximum likelihood estimation of optimal scaling factors for expression array normalization. *SPIE BIOS* 2001.

Hochreiter S, Clevert DA and Obermayer K (2006). A new summarization method for affymetrix probe level data. *Bioinformatics* 22(8):943–949.

Hoffmann R, Seidl T, Dugas M. (2002). Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. *Genome Biology* 3(7).

Huber W., Von Heydebreck A., Sültmann H., Poustka A. and Vingron M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 18: suppl. 1 (2002), S96-S104 (ISMB 2002).

Irizarray RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* 31:e15.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. and Speed, T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4(2):249-64.

Kong, Y. M., Chen, Z. , Cai, J., and Scheuermann, R. (2007). Use of Gene Ontology as a Tool for Assessment of Analytical Algorithms with Real Data Sets: Impact of Revised Affymetrix CDF Annotation. 7th International Workshop on Data Mining in *Bioinformatics (BIOKDD* 2007), page 64-72.

Kong, Y. M., Chen, Z., Qian, Y. McClellan, E., McGee, M. and Scheuermann, R.H. (2008). Objective selection of the optimal microarray analysis pipeline. In preparation.

Lee, J.A., Sinkovits, R.S., Mock, D., Rab, E.L., Cai, J., Yang, P., Saunders, B., Hsueh, R.C., Choi, S., Subramaniam, S., Scheuermann, R.H. in collaboration with the Alliance for Cellular Signaling. (2006). Components of the antigen processing and presentation pathway revealed by gene expression microarray analysis following B cell antigen receptor (BCR) stimulation. *BMC Bioinformatics* 7: 237.

Li C, Wong HW. (2001a). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences*. 98:31-36.

Li C, Wong HW. (2001b). Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology*. 2:research0032.1-0032.11.

McGee., M. and Chen, Z. (2006a). New spike-in data sets for the Affymetrix HG-U133a Latin square experiment. *COBRA* Reprint Series, Article 5.

McGee M and Chen Z (2006b). Parameter Estimation for the Exponential-Normal Convolution Model for Background Correction of Affymetrix GeneChip Data. *Statistical Applications in Genetics and Molecular Biology* 5, Article 24.

McGee M, Chen Z, Luo F, and Scheuermann RH (2006). A Distribution-Free Convolution Model for Background Correction of Oligonucleotide Microarray Data. SMU Technical Report 340, http://www.smu.edu/statistics/TechReports/tech-rpts.asp

Mosteller F and Tukey J (1977). Data Analysis and Regression. Reading, MA: Addison-Wesley.

Quackenbush J (2002). Microarray data normalization and transformation. *Nature Genetics* 32, 496-501.

Schuchhardt J, Beule D, Malik A, Wolski E, Eickhoff H, Lehrach H and Herzel, H (2000). Normalization Strategies for cDNA microarrays. *Nucleic Acids Res.* 28 (10): e47.

Shi, L., Reid, L.H., Jones, W.D., Shippy, R., Warrington, J.A., Baker, SC., Collins, P.J., de Longueville, F., Kawasaki, E.S., Lee, K.Y., Luo, Y. et al. (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol.* 24(9):1151-1161.

Therneau T and Ballman KV (2005). What does PLIER really do? Technical Report 75, The Mayo Foundation.

Warren, P., et al. (2007). "PANP - a New Method of Gene Detection on Oligonucleotide Expression Arrays," *Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering*, 108-115.

Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F. (2004). A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association*. 99:909-917.

Wu, Z. and Irizarry, R.A. (2005). "A Statistical Framework for the Analysis of Microarray Probe-Level Data," *Johns Hopkins University, Dept. of Biostatistics Working Papers.* Working Paper 73 http://www.bepress.com/jhubiostat/paper73.